FEMS Microbiology Letters, 2025, 372, fnaf051
DOI: 10.1093/femsle/fnaf051
Advance access publication date: 22 May 2025
Research Article – Taxonomy, Systematics and Evolutionary Microbiology

Surprising effects of differential loss in genome evolution: the last-one-out

Nico Bremer [1,*], William F. Martin[1], Mike Steel[2]

[1]Faculty of Mathematics and Natural Sciences, Institute of Molecular Evolution, Heinrich Heine University Düsseldorf, 40225 Duesseldorf, Germany
[2]Biomathematics Research Centre, University of Canterbury, 8140 Christchurch, New Zealand
*Corresponding author. Faculty of Mathematics and Natural Sciences, Institute of Molecular Evolution, Heinrich Heine University of Duesseldorf, 40225 Duesseldorf, Germany. E-mail: nico.bremer@hhu.de
Editor: [Aharon Oren]

Abstract

Gene loss is an important process in genome evolution, though its power is often underestimated. If a gene is present at the root of a phylogenetic tree and can be lost in one lineage across the tree, it can potentially be lost in all, leading to gene extinction. Just before gene extinction, there will be one lineage that still retains the gene, generating a "last-one-out" distribution. Such an isolated gene presence will emulate the result of recent lateral gene acquisition, even though its distribution was generated by loss. How probable is it to observe "last-one-out" distributions in real data? Here, we mathematically derive this probability and find that it is surprisingly high, depending upon the tree and the gene loss rate. Examples from real data show that loss can readily account for observed frequencies of last-one-out gene distributions that might otherwise be attributed to lateral gene transfer.

Keywords: gene loss; lateral gene transfer; birth–death process

Introduction

Gene loss is an important and ubiquitous mechanism of genome evolution. In prokaryotes, gene loss acting on the whole genome is traditionally called reductive evolution (Andersson and Kurland 1998, van Ham et al. 2003, Oshima et al. 2004, Hosokawa et al. 2006) and can result in miniscule genome sizes in parasites and endosymbiotic bacteria, the current record being Macrosteles quadrilineatus (Moran and Bennett 2014) an endosymbiotic bacterium of leafhoppers that harbors only 137 protein-coding genes. Reductive evolution is also observed in symbiotic archaea (Waters et al. 2003) and in eukaryotes, especially among intracellular parasites (Tovar et al. 2003, Nicholson et al. 2022). Genome reduction through gene loss is also the central underlying theme of genome evolution in mitochondria and plastids, the endosymbiotic organelles of eukaryotic cells (Moore and Archibald 2009), which can sometimes lose their genomes altogether (Müller et al. 2012), because many genes lost from organelle genomes have been transferred to the nucleus (Martin et al. 1998, Timmis et al. 2004). In eukaryotes, gene loss is also very common and particularly well studied following whole-genome duplications (Blanc and Wolfe 2004, Kellis et al. 2004, Brunet et al. 2006, Scannel et al. 2006), where duplicate gene copies are rapidly lost by mutation, restoring diploid genetics in chromosome polyploids (Blanc and Wolfe 2004). Additionally, gene loss is often seen as a driving factor in genome evolution (Olson 1999, Albalat and Cañestro 2016, Guijarro-Clarke et al. 2020). In general, if a gene belonging to a clade can be lost once in one lineage during evolution, it can be lost again in other lineages as well.

In comparative genome studies, gene loss is easy to detect if losses are rare, as shown in Fig. 1. If most genomes in a sample contain a given gene of interest, but one or a few do not, there can be little doubt that gene loss has occurred in the genomes lacking the gene. But the more common loss is, the more difficult it becomes to distinguish from lateral gene transfer (LGT). If a given gene is present in about half of the genomes in a sample, the decision between loss and LGT becomes a matter of weighing the relative probabilities of LGT and gene loss, entailing an *a priori* assumption that LGT is roughly as common as loss. In eukaryotes, gene loss is much more common than LGT from prokaryotes (Ku et al. 2015, Ku and Martin 2016). But there have been a number of highly publicized claims for widespread LGT to eukaryotes, though, for example "hundreds" of LGTs in the human genome (Consortium 2001) or fully "17%" of the tardigrade (a primitive animal) genome being the result of recent LGTs (Boothby et al. 2015). Both the human genome LGT claims and the tardigrade LGT claims were reinspected and turned out to be data contaminations and data interpretation problems, not LGT (Salzberg et al. 2001, Stanhope et al. 2001, Koutsovoulos et al. 2016, Salzberg 2017). Most reports of LGT in eukaryotes are not critically reinspected, but they are highly cited (Martin 2017, Keeling 2024). Nonetheless, there are cases where contaminations can effectively be ruled out—for example when a gene in question is observed in the genome sequence of several individuals from a given species (Koutsovoulos et al. 2016). Distributions of the type seen in Fig. 1(c) are observed in real data for eukaryote genomes, where LGT might appear to be the most likely cause, because the possibility of many independent losses as the cause of the pattern would seem, at face value, extremely unlikely.

But is gene loss leading to last-one-out topologies really unlikely? Or do we just assume it is unlikely, thereby opting to sug-

Received 4 March 2025; **revised** 29 April 2025; **accepted** 20 May 2025

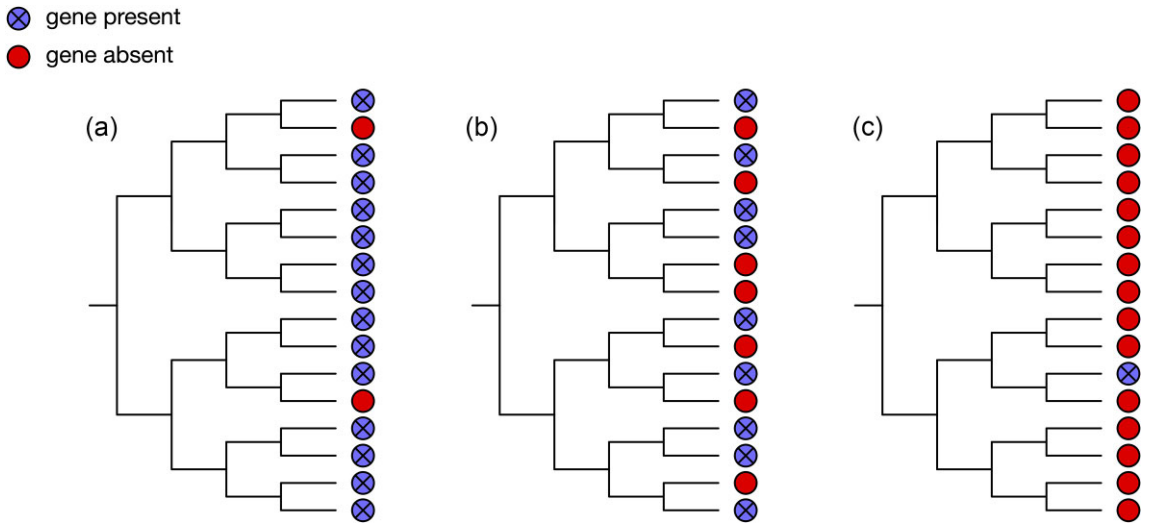

Figure 1. Hypothetical phylogenetic species trees showing the presence and absence of genes across all species in the trees. A circle with a cross indicates that the gene is present in this species, an empty circle indicates that a gene is absent. (a) A distribution where gene loss most likely appeared on the branches to two species. (b) A case where the distribution of the genes that are present and absent is almost equal across the species tree. The decision between LGT and gene loss is highly dependent on the weighing of their relative probabilities. (c) Illustrates a case where the gene is only present in one species. An easy (but not necessarily true) explanation for this would be LGT. This gene distribution across the tree can also be the result of a minimum of four gene losses if the gene was already present at the root node.

gest that LGT was at work without even testing the possibility that the distribution is actually the result of many independent losses. Are there even tools available to test such a case? This answer, until now, has been no. Many analytical tools to study prokaryotic genomes are currently in use that employ different and usually predetermined gain/loss ratios that are designed to differentiate between loss and LGT (Goodman et al. 1979, Page 1994, Bansal et al. 2012, Szöllősi et al. 2013). In many cases, the overall average ratio of gene loss to LGT ends up being close to 1 in such applications, for obvious reasons. If loss predominates, then genomes steadily decrease in size across the reference tree (that is, ancestral genomes inflate), and if LGT predominates, genomes steadily increase in size across the reference tree (that is, ancestral genomes become too small) (Dagan and Martin 2007). Some tools for estimating loss versus LGT in current use can entail differences in loss versus transfer probabilities for individual genes that differ by 20 orders of magnitude (Bremer et al. 2022).

If gene loss is the predominant mode of genome evolution for a given gene in a given group, it will become lost in many lineages, ultimately in all. Just before the gene goes extinct in the group, however, there will exist a state in which the gene is present in only a few genomes, and finally, over time, only in one genome of the group. If this gene is in a eukaryote, but has homologs in prokaryotes, gene loss in eukaryotes will produce a pattern that looks exactly like LGT: The gene is present in prokaryotes and one (or a few) eukaryotes. Under a loss-only mode of evolution, the last-one-out looks like an LGT, but the pattern was generated solely through gene loss. Here, we address the question of how likely it is to observe a last-one-out gene distribution under loss-only models.

Results

Mathematical modeling and algorithms

We now describe mathematical and computational methods to investigate the probability of last-one-out scenarios in both synthetic and real trees. We assume that each gene in a phylogeny

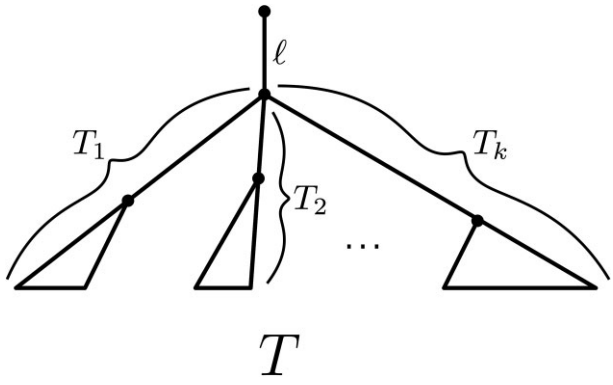

Figure 2. Hypothetical phylogenetic tree T with the subtrees $T_1, T_2 \ldots T_k$ and branch length ℓ.

can be lost along each lineage of a tree according to a continuous-time Markov process with loss rate μ, and which operates independently across genes and lineages.

Recursion for a given tree

Let T be a rooted tree with a stem edge of length ℓ, and let $T_1, T_2 \ldots T_k$ denote the subtrees of T incident with this stem edge, as shown in Fig. 2. Although the lengths of edges may correspond to time, and so be ultrametric, the algorithm described in this first section does not assume that edge lengths are ultrametric. Let π_T^+ denote the probability that a gene g that is present at start of the stem edge of T is present in *exactly one* leaf of T, and let π_i^+ denote $\pi_{T_i}^+$ (the corresponding probabilities for the subtrees $T_1 \ldots, T_k$). To calculate π_T^+ recursively, we also need to calculate the probability π_T that g is not present at any of the leaves of T, and we let π_i denote π_{T_i}.

Note that if T consists of just a single stem edge of length ℓ (the base case in the recursion), then $\pi_T = 1 - e^{-\mu\ell}$ and $\pi_T^+ = e^{-\mu\ell}$. Thus we may suppose that $k \geq 2$. The following result (proved in the Appendix) provides a polynomial-time way to compute these quantities recursively via dynamic programming (progressing from the

leaves to the root). Note that both Parts (i) and (ii) are required for computing π_T^+.

Proposition 1. For the tree shown in Fig. 2, the following recursions hold:

$$\pi_T = \left(1 - e^{-\mu\ell}\right) + e^{-\mu\ell}\pi_1\pi_2\ldots\pi_k, \tag{1}$$

$$\pi_T^+ = e^{-\mu\ell}\pi_1\pi_2\ldots\pi_k\left(\frac{\pi_1^+}{\pi_1} + \frac{\pi_2^+}{\pi_2} + \ldots + \frac{\pi_k^+}{\pi_k}\right). \tag{2}$$

For binary trees, Equation (2) simplifies to:

$$\pi_T^+ = e^{-\mu\ell}\left(\pi_1\pi_2^+ + \pi_1^+\pi_2\right). \tag{3}$$

If there are $G \geq 1$ genes present at the top of the stem edge of T, and losses occur independently among the genes (each with rate μ), then the number of genes that appear in just one leaf of T has a binomial distribution with parameters $(G,\ \pi_T^+)$.

To illustrate Proposition 1 with a simple example, consider the tree in Fig. 2, where each of the subtrees T_1..., T_k is a single leaf at the same distance from the root, and $\ell = 0$ (the "star tree"). Under the gene-loss model, a gene that is present at the root of the tree will be present at exactly one leaf of this tree precisely if there are *exactly* $k - 1$ loss events. This might seem very unlikely for large values of k. However, the probability of this event can be as large as $e^{-1} = 0.367$ even as k becomes large, provided that μ is chosen appropriately (and dependent on k); details are provided in the "Analysis of the star tree" section of the Appendix. Nevertheless, if we consider the posterior value of this probability by taking a uniform prior on $1 - e^{-\mu}$ (setting the height of the tree to 1), then this posterior probability tends to 0 as the number of leaves of the tree (k) grows. The proof of these claims and the analysis of this star tree when we allow $\ell > 0$ are provided in the Appendix. Of course, the star tree is a highly nonbinary tree, which raises the question of whether π_T^+ can be close to e^{-1} when T is binary and the number of leaves is large. This is indeed possible: we can simply resolve the polytomy at the root by using very short interior edges to obtain a binary tree for which π_T^+ will be close to the corresponding value for a star tree and hence can be close to e^{-1} for a suitably chosen value of μ. However, for trees generated by simple phylodynamic models, this is no longer the case, as we demonstrate in the next section. The analyses in this manuscript mainly focus on binary trees.

Random trees

Suppose now that T is generated by a standard birth–death model (Kendall 1948, Lambert and Stadler 2013) with speciation rate λ and extinction rate ν, starting from a single lineage at time t in the past. The tree T is now a random variable, denoted T_t, and the number of species at the present (denoted N_t) is also a random variable and has a (modified) geometric distribution with expected value $\mathbb{E}[N_t] = e^{(\lambda-\nu)t}$. We will suppose that $\lambda > \nu$ since otherwise the tree T_t is guaranteed to die out as t grows. Let π_t^+ be the probability that a gene g that is present at start of the stem edge of T_t is present in *exactly one* leaf of T_t. The following result precisely describes the maximum value that π_t^+ can take as μ (the rate of gene loss) varies over all possible positive values. The short proof is provided in the Appendix.

Proposition 2.

$$\max_{\mu}\pi_t^+ = \frac{1}{(1+\lambda t)^2} = \frac{1}{\left(1 + \frac{\ln \mathbb{E}[N_t]}{1-\frac{\nu}{\lambda}}\right)^2}. \tag{4}$$

Notice that although $max_\mu \pi_t^+ \to 0$ for Yule trees as they grow in their expected size, the convergence is quite slow as a function of the expected number of leaves of the tree, due to the presence of the logarithmic function on the right of Equation (4). Also, if there are $G \geq 1$ genes present at time 0, then the expected number of genes that will be present in just leaf of T_t is $G \cdot \pi_t^+$. However, in contrast to Proposition 1(iii), the number of genes present in just one leaf of T_t is no longer binomially distributed, since this number is now a compound random variable because it is dependent on the random variable T_t.

To illustrate Proposition 2, consider (pure–birth) Yule trees (i.e. $\nu = 0$) with an expected number of 150 leaves. Then $max_\mu \pi_t^+ \approx 0.028$, and so for 10 000 independent genes and this optimal rate of gene loss, the expected number of genes that would be last-one-out (i.e. present in just one leaf of these Yule trees) would be around 280. This provides some insight into the results described in the next section.

Application to real genome data

To test this algorithm on real genome data we chose the example of genes in eukaryotic genomes that have homologs in prokaryotes but that are present in only one or a few eukaryotic lineages. Such patterns are taken as evidence for the workings of differential loss, under the assumption that loss will generate such patterns (Ku et al. 2015), or as evidence for the workings of LGT (Cote-L'Heureux et al. 2022) under the assumption that LGT rather than loss generates such patterns. The calculation of the probability of a gene being present at the root node and remaining in exactly one leaf of a eukaryotic tree requires a rooted species tree and a gene loss rate μ. Reconstructing a eukaryotic species tree is challenging, and there is currently no consensus on the position of the root (Keeling and Burki 2019, Burki et al. 2020). Although the loss rates can be adjusted and averaged across a range of values, the backbone trees with all their nodes, branches and branch lengths are not that easily adjustable.

We started by investigating a set of ten eukaryotic gene trees with 150 leaves each. These gene trees need not be representative of the true phylogeny of eukaryotes, nor need they show a pattern of gene distribution that could be indicated as LGT. They are just used to test the algorithm, whereby the different trees were selected merely to show that different phylogenies can have an influence on the calculated probability of a gene being present at the root node and remaining only in one leaf of a eukaryotic tree. Furthermore, the different gene trees with 150 leaves provide an opportunity to estimate the overall probability of observing a last-one-out pattern if we consider thousands of eukaryotic genes with prokaryotic homologs (Fig. 3). In Fig. 3, we assume that 10 000 genes were present in the last common ancestor of 150 eukaryotes. For those trees, the mean probabilities across a range of different loss rates μ, where $1 - e^{-\mu}$ ranges from 0 to 1, result in 232 (lowest mean) to 790 (highest mean) last-one-out cases that would look like LGT but actually are the result of differential loss in a loss-only mode of evolution for a 10 000 gene ancestral genome. Looking at the median, we would find 11 (lowest median) to 755 (highest median) cases, depending on the tree itself. Since loss rates are not constant over time, we cannot assume that these percentages resemble the "real" amount of those cases due to differential loss. This first look into the data with our new tool does show, however, that last-one-out cases are by no means so rare that they can be excluded *a priori*. If the loss rate is ideal, meaning that the maximum probability of last-one-out cases for the given tree is achieved, we would see between 532 (lowest maximum) and 2502 (highest maximum) out of the 10 000 genes resulting in a last-one-out scenario, which is a substantial frequency. That is, in a study of 10 000 gene families present in the eukaryotic common

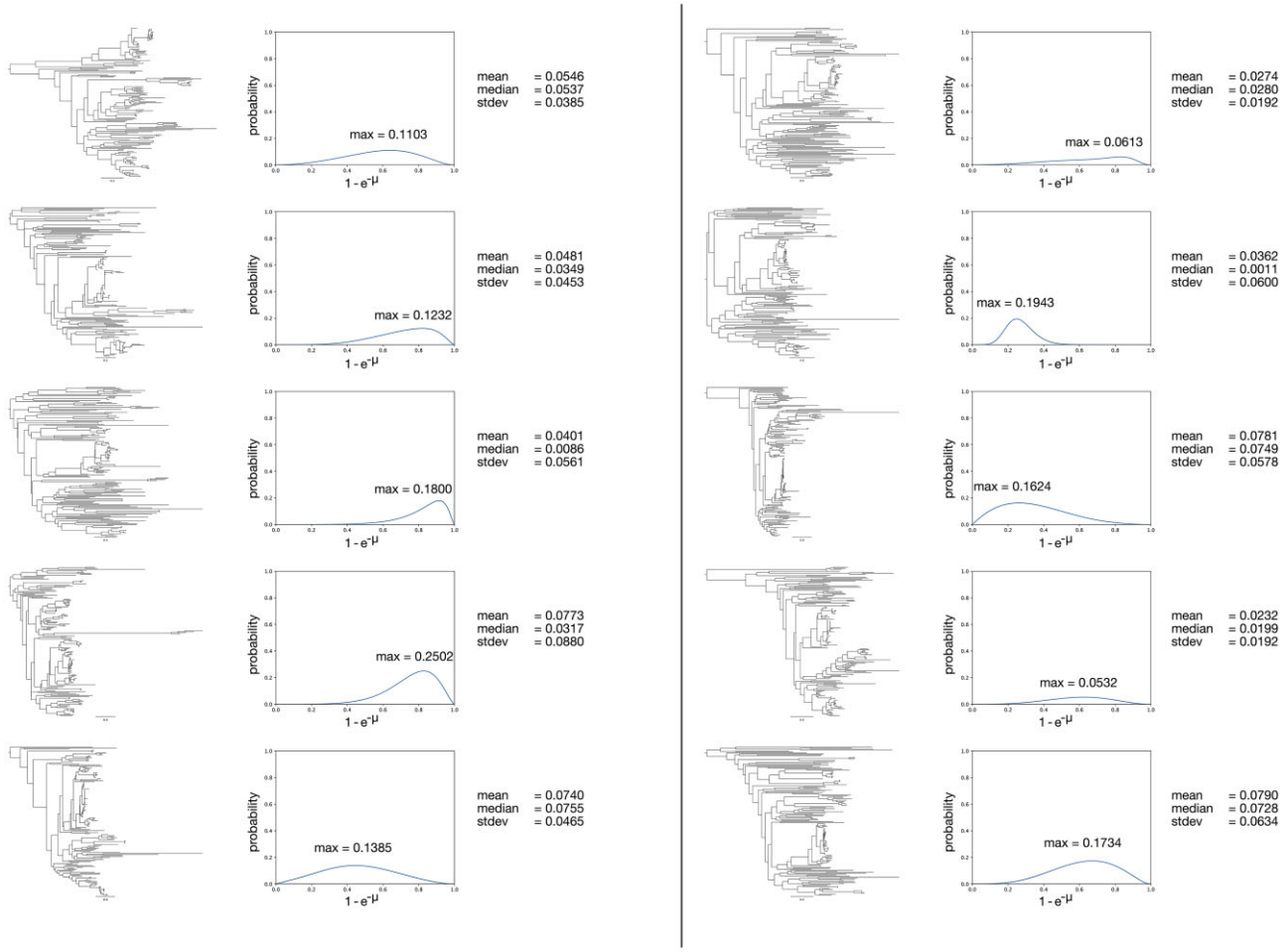

Figure 3. Ten eukaryotic gene tree phylogenies with 150 leaves each and the corresponding probabilities for a last-one-out scenario against $1 - e^{-\mu}$ (μ = gene loss rate). The trees show various possibilities of species trees without assuming that those trees represent a real eukaryotic backbone tree. They show that the phylogeny itself has an influence on the probability of a last-one-out scenario, but that the overall probability is comparably high.

ancestor, one would expect to observe dozens, hundreds, or even thousands of last-one-out patterns in trees sampling 150 genomes obtained solely as the result of differential loss. Put another way, we would expect to observe last-one- out distributions at a frequency that is not far off from the number of genomes in the tree. These cases would appear, in a gene phylogeny, as a single eukaryote (or group thereof) branching within prokaryotic homologs. Such cases are observed with real data.

Three test cases

The surprisingly high probability to observe a gene that is present in the root node and only in one species or clade and lost in all other leaves of a tree offers a new approach to investigate data that looks like evidence for LGT based on a rare or sparse gene distribution. Differential loss can—and will—produce last-one-out patterns that look just like lineage specific LGT. It is therefore possible, if not probable, that some reports suggesting evidence for LGT are perhaps last-one-out cases attributable to differential loss. We asked whether we could identify such cases in real data. In the following, we examine two studies that propose LGT as the cause of last-one-out topologies. Our aim is not to challenge these specific papers, but simply to see if the model proposed here (requiring only a tree, gene loss, and a specific case of gene distribution) can account for the data directly, without recourse to LGT. The aim of our study was to get an exact method to calculate the probability of observing such an event under a loss-only process.

One recent study is very helpful. Cote-L'Heureux et al. (2022) looked for lineage-specific presence of prokaryotic genes in eukaryotes that would provide the strongest possible evidence, in their view, for the workings of LGT from prokaryotes to eukaryotes. They sampled 13 600 gene families, 189 eukaryotic genomes and 540 eukaryotic transcriptomes, looking for recent lineage-specific LGT (topologies that we designate as last-one-out patterns). Among the 13 600 eukaryotic gene families sampled, they found ~94 putative cases of LGT that represent a last-one-out pattern, that is, a restricted single-tip distribution of a prokaryotic gene in a eukaryotic genome or group, which they interpreted as strong evidence for LGT. Our present findings (Fig. 3) indicate that in Cote-L'Heureux et al. (2022) the number of cases identified in their study (94) is very close to the lower bound of the expectations for last-one-out topologies of similarly sized data sets, in which all the last-one-out topologies can be accounted for by differential loss alone, with no need to invoke LGT.

One clear prediction of lineage-specific LGT versus loss for last-one-out cases is this: If lineage-specific acquisition is the mechanism behind the observed rare presence pattern for a eukaryotic gene, then the acquisition would need to be evolutionarily late (that is, a tip acquisition). That is, the prokaryotic donor and the eukaryotic gene should share a higher degree of sequence

similarity, on average, in comparison to genes that trace back to the eukaryotic common ancestor. This is the reasoning behind the analysis of Ku et al. (2015) and Ku and Martin (2016), who looked for evidence of recent acquisitions of prokaryotic genes in sequenced eukaryotic genomes. Ku et al. (2015) found that, in eukaryotic genomes, rare genes that have prokaryotic homologs were not more recently acquired (that is, they were not more similar to prokaryotic homologs) than genes that trace back to the eukaryotic common ancestor, suggesting that their rare occurrence is the result of differential loss rather than lineage-specific acquisition (Ku et al. 2015, Ku and Martin 2016) (Fig. 4a and b).

Cote-L'Heureux et al. (2022) employed the same test, making the same kind of comparison that Ku et al. (2015) performed, namely, they looked for cases in which the prokaryotic gene was acquired recently by the eukaryotic lineage, using the criterion of sequence similarity. What they found was the distribution shown in Fig. 4(c), namely that the cases they suspected to be LGTs were just as old, in terms of sequence divergence, as genes that were acquired from the mitochondrion. In other words, there were no obviously recent acquisitions, as all of the prokaryotic genes that they interpreted as recent LGTs had the hallmark of ancient acquisition, just as Ku et al. (2015) suggested. Cote-L'Heureux et al. (2022) offered no explanation for the finding that genes they interpreted as recent acquisitions via LGT were just as ancient, in terms of sequence identity, as genes acquired from mitochondria (Fig. 4c). One interpretation is that the genes in their LGT class were not LGTs after all but were the result of differential loss instead. Differential loss directly explains why such genes show just as much sequence divergence to prokaryotic homologues (Ku et al. 2015) (Fig. 4c) as genes present in the eukaryotic common ancestor. LGT models would need to invoke an *ad hoc* corollary assumption of substitution rate acceleration for every gene with a last-one-out pattern to account for the absence (Fig. 4c) of eukaryotic LGTs having high (> 70%) sequence similarity to prokaryotic homologs. Differential loss requires no rate acceleration corollary. Furthermore, the model presented here closely predicts the frequency of observing last-one-out patterns under a variety of topologies and loss rates, which becomes increasingly relevant as new data point to a gene rich mitochondrial ancestor (Leger and Gawryluk 2024).

A second recent study provide an additional opportunity to test the method. We investigated a dataset of 332 budding yeast species published by Shen et al. (2018). They reported 365 distinct events of horizontal gene transfer to yeast lineages. Of those, 230 appeared to be species-specific. Could those 230 cases also be the result of differential loss, unsuspected "last-one-out" cases? We analyzed the species tree provided in Shen et al. (2018) for "last-one-out" probabilities. The authors reported that the last common ancestor of budding yeasts was similar to an archaetypal member of its sister subphylum Pezizomycetes with ∼10 000–13 000 genes. We were therefore able to calculate the mean and median probabilities for a "last-one-out" and therefore the expected number of these events (Fig. 5). The mean probability across all possible loss rates in the interval between 0 and 1 for $1 - e^{-\mu}$ is roughly 0.0287 and the median probability for this interval is 0.0229. If the last common ancestor of budding yeasts had 10 000–13 000 genes, this results in 287–373 genes being statistically the "last-one-out" for the mean probabilities and 229–298 "last-one-out" cases for the median probability. In comparison to the 230 analyzed cases that are supposed to be the result of LGT according to Shen et al. (2018), the statistical probabilities of observing these "last-one-out" cases though differential loss (rather than LGT) is not unlikely at all, it is in agreement with the expectation, and our simple model is surprisingly accurate.

As a third example, we examined a dataset where differential loss rather than LGT was reported to be the cause of last-one-out topologies. Ku et al. (2015) clustered 956 053 protein sequences from 55 eukaryotic genomes across six supergroups and compared them to a total of 6 103 025 protein sequences from prokaryotes across 1847 bacterial and 134 archaeal genomes. They found a total of 2585 eukaryote–prokaryote clusters and 101 eukaryotic singletons [Supplemental Table 9 in Ku et al. (2015)] with prokaryotic homologs. The ancestral genome size of the last eukaryotic common ancestor of this dataset comprises 2686 genes. The calculation of last-one-out probabilities yielded a mean probability of 0.0967 and a median probability of 0.1057 (Fig. 6). Since the ancestral genome size was 2686 genes for this dataset, one would expect on average 260 last-one-out cases and 284 cases using the median probability using our method; the 101 last-one-out cases observed are fewer than expected. In this example, our simple algorithm again works on real world data. Considering that this algorithm is based on a loss-only model of evolution, the lower number of observed cases compared to expected cases will likely be the result of gene duplications, which play a significant role in eukaryotic evolution (Scannel et al. 2006, Hittinger and Carrol 2007, van de Peer et al. 2009), within this eukaryotic data set.

The role of selection

How do gene duplications and selection figure into this issue? Gene duplications, genome duplications, and recurrent duplications leading to gene family expansions lead to growth of gene families during evolution. Members of such families can, and do, undergo differential loss in different lineages. For a gene that was present in the eukaryote common ancestor, and that underwent loss across eukaryotic lineages in such a manner as to generate a last-one-out topology, the question arises about the role of selection in that process. Clearly, if the gene in question was essential during eukaryote evolution, selection would have maintained its presence in all lineages. Losses indicate phases of evolution in which the gene was required under some conditions for some lineages, with relaxed or absent selective pressures in others, allowing loss in some lineages but retention in others, possibly as a result of persistent selection or the gene having acquired a novel function. Such an example can be found in the evolution of Fe–Fe hydrogenases, where the gene is present and functional among green algal lineages, which often experience anaerobiosis (Happe and Kaminski 2002). Yet, during the transition to life on land in an atmosphere of 21% O_2, the gene lost its function in anaerobic energy metabolism, whereby a duplicate of the Fe–Fe hydrogenase gene acquired a new function in O_2-sensing in the land plant lineage instead (Gould et al. 2019). Changing environments or developmental contexts can alter the selective pressures that act upon gene retention, gene loss, or gain-of-function (Li et al. 2019).

Of course, it is also possible that selective pressures could, in principle, lead to gene gain via LGT. What form of selection is strong enough to cause LGT on time frames, where we can directly observe the effects? The experiment has already been done, we just need to tally the result—growth inhibitors. The best known, and best studied, example of selection for LGT is the spread of antibiotic resistance genes across bacteria in hospitals starting in the 1950s, which led to the discovery of both plasmids and LGT among bacteria. A literature search returns, for example, over 97 000 papers on bacteria* and antibiotic* and resistance* with over 14 000 of those papers containing the search terms plasmid* or transfer

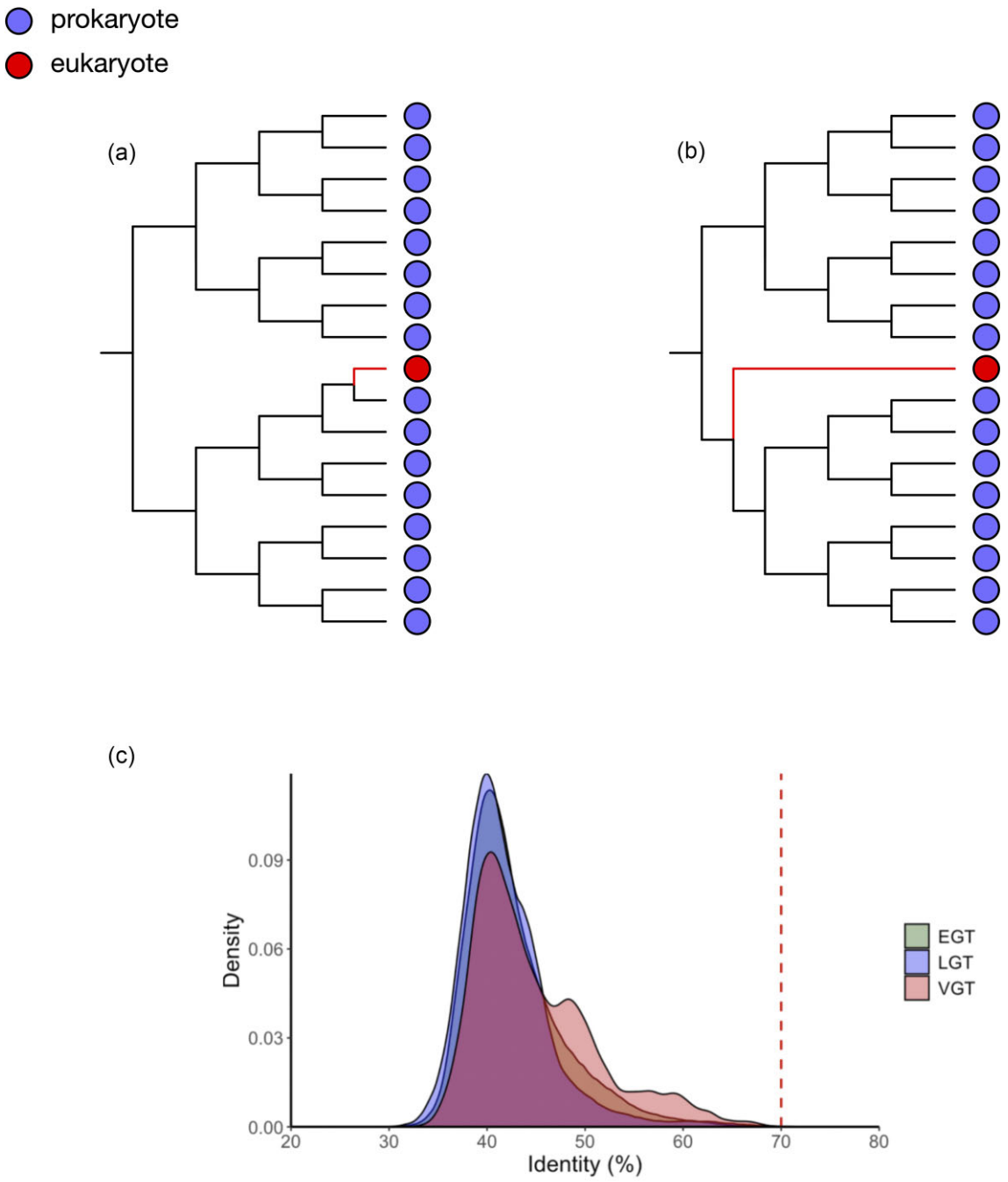

Figure 4. Similarity of eukaryotic last-one-out cases to prokaryotic homologs. (a) Phylogenetic distribution of genes, where the eukaryotic gene is considered to be the result of LGT due to its high similarity to one prokaryotic homolog. (b) The eukaryotic gene does not have a substantially high similarity to its prokaryotic homologs. It can therefore not be the result of recent LGT and is more likely the result of differential gene loss. (c) Supplementary Figure 9 from Cote-L'Heureux et al. (2022) showing that genes assumed to be the result of LGT are at most 70% similar to their prokaryotic homologs. This finding supports the "70% rule" of Ku and Martin (2016) and furthermore shows that these cases are more likely to be the result of differential loss instead of LGT. EGT: endosymbiotic gene transfer (genes acquired from chloroplasts or mitochondria); LGT: lateral gene transfer; and VGT: vertical gene transmission.

(The "∗" within the literature search are wildcards that stand for any amount of possible characters). Though a small sample, this underscores a point well-known among all microbiologists: bacteria immediately respond to antibiotic selective pressure by acquiring resistance genes via LGT. Antibiotic resistance represents a clear case of gene gain bringing selective benefit, in bacteria. Is the same true for eukaryotes, where reports for LGT have recently been reviewed (Keeling 2024)?

The closest eukaryotic equivalent to antibiotics in hospitals would be the use of fungicides in agriculture, which have been in use for over a century (Russell 2005). Does the strong selective pressure exerted by fungicides also lead to LGT among fungi, a group for which reports of LGT have also been reviewed (Richards and Talbot 2013)? A literature search returns over 13 000 papers on fung∗ and fungicid∗ and resistance∗, yet only 12 of those papers contained the search terms "horizontal gene transf∗" or "lat-

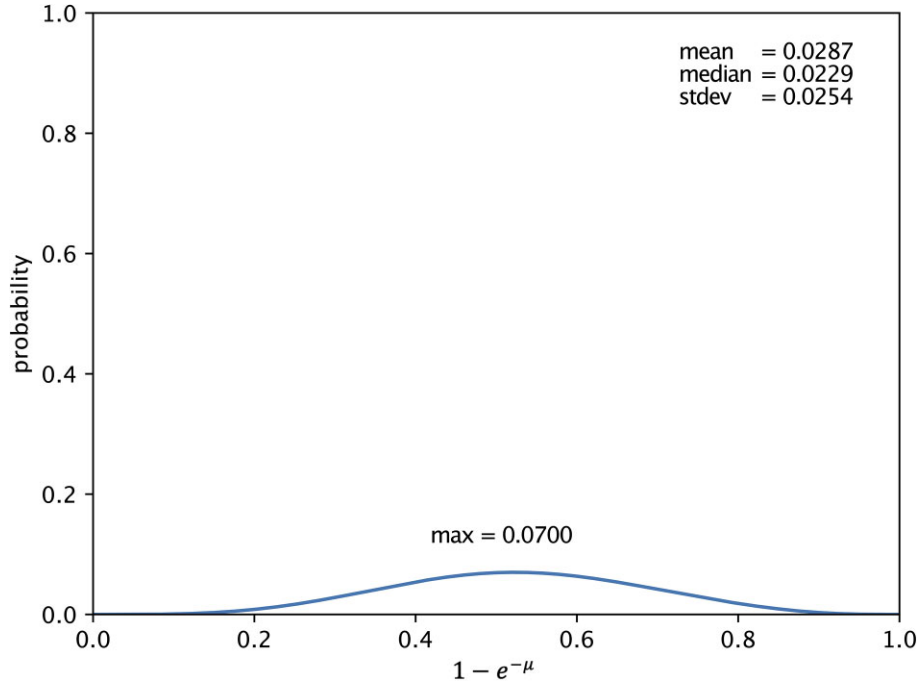

Figure 5. Probabilities of last-one-out cases across a spectrum of loss rates.

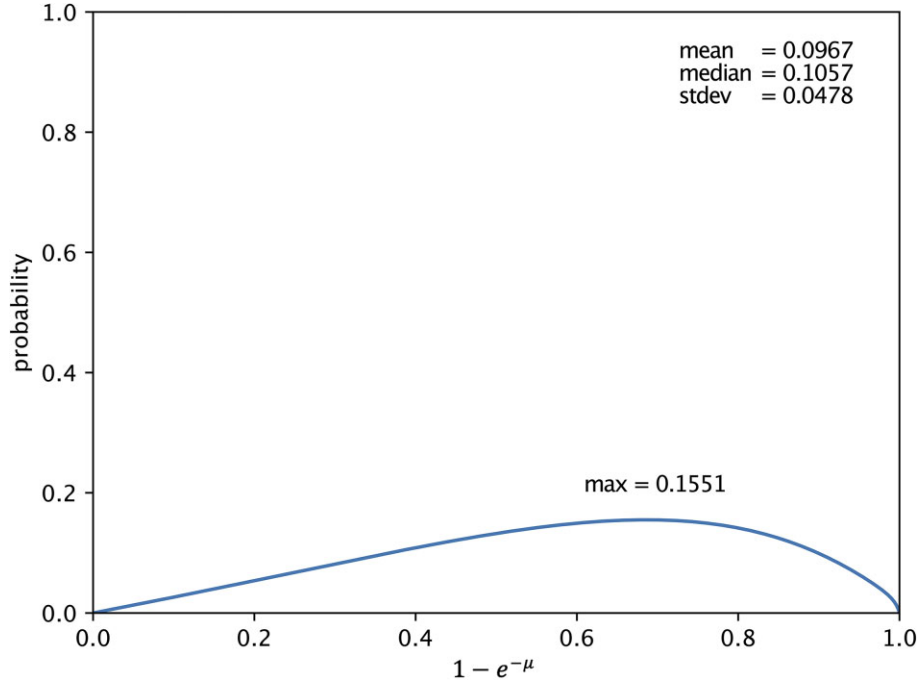

Figure 6. Probabilities of last-one-out phylogenies across a spectrum of loss rates for a eukaryotic species tree of 55 genomes from Ku et al. (2015) with forced monophyly for eukaryotic groups using a concatenated alignment of five genes universally present in those 55 genomes. For species that were originally used in Ku et al. (2015) but do not have a representative in the RefSeq dataset, we chose alternative genomes that are taxonomically near the original genomes. Alignments were generated using Multiple Alignment using Fast Fourier Transform (MAFFT) (Katoh et al. 2002), using the iterative refinement method that assimilates local pairwise alignment information (L-INS-i). The tree was constructed with IQ-Tree (Nguyen et al. 2014), using the best-fit model and forcing monophyly of eukaryotic groups described in Ku et al. (2015) and the tree was rooted with MAD (Tria et al. 2017).

eral gene transf∗"and only one of those 12 papers reported a case of phylogenetic evidence for LGT among fungi for a putative (not documented) resistance gene against cyanate (Elmore et al. 2015). The other 11 papers were mainly about fungicide-induced mobilization of plasmids in bacteria. Fungicide resistance represents a clear case where gene gain via LGT could bring (life-saving) benefit against a lethal selective pressure, but LGT is not observed. The relative frequency of LGT events conferring resistance to growth inhibitors (14 000 reported cases for bacteria, one possible candidate case reported for fungi) suggests that bacteria and fungi respond very differently to selection pressure generated by growth inhibitors. How do fungi respond?

Of course, resistance to fungicides is widespread and well-known in agriculture, as are resistance to insecticides and herbicides. Yet, the many known cases of resistance against fungicides, insecticides and herbicides do not entail LGT; instead, they involve *de novo* point mutations in the target-site encoding genes (fungicides), selection of polygenic metabolic resistance from standing variation (herbicides), and a combination of standing variation and *de novo* mutations in the target site or major metabolic resistance genes (insecticides) (Hawkins et al. 2019). That is, humans have already performed the experiment involving the application of strong selection pressure to prokaryotes (antibiotics in hospitals) and to eukaryotes (agricultural pests), and the result is generally clear. Bacteria respond by LGT of preexisting resistance genes, while eukaryotes respond by *de novo* point mutations and sexual recombination of standing variation, at least in cases reported so far. The present comparative example from the application of strong selective pressure indicates that there is a fundamental difference between prokaryotes and eukaryotes with regard to their tendency to acquire genes via LGT in response to strong (lethal) selection. However, both prokaryotes and eukaryotes do undergo gene loss at high rates.

As a caveat, strong selection for resistance is just one ecological context. A literature search for the search terms "fung∗" and "horizontal gene transf∗" or "lateral gene transf∗" returns over 1000 papers on fungal LGT (though none for resistance against fungicides), and we are by no means suggesting that those 1000 papers, usually founded in genome comparisons and gene phylogenies, and many invoking trait selection, are in error, collectively or otherwise. However, a recent study reexamined the strength of phylogenetic claims for LGT among fungi and found that only about 1.5% of trees that have been published as evidence for LGT among ascomycetes (the group of fungi that includes yeast) withstand critical inspection (Aguirre-Carvajal et al. 2025). If one accepts phylogenetic evidence for LGT among eukaryotes, the lack of abundant reports indicating eukaryotes to respond to strong selective pressure with LGT of resistance genes, while prokaryotes obviously do respond to selection with LGT, presents a puzzling observation. The ability of gene loss to generate last-one-out topologies at surprisingly high frequencies, as we have demonstrated here, might help to reconcile some discrepancies and contribute to solving the puzzle.

Conclusion

Sparse gene distributions in eukaryotes are often interpreted as evidence for gene acquisition via LGT from prokaryotes. However, gene loss can generate the same patterns, but estimates for the probability of observing a single gene at the tip of a phylogenetic tree as the result of differential loss within a given clade, as opposed to LGT, have been lacking, because methods were not even available. Here, we have derived the probability of observing such cases, which we call last-one-out patterns, because under a loss-only model, the last gene to be lost looks like an instance of LGT. The probability depends on the size and shape of the tree, and the loss rate μ. We find that the probability of observing a last-one-out topology can be (surprisingly) high.

This is not to say that there is no LGT to eukaryotes at all. But if LGT to eukaryotes were as common as many reviews would have us think, there have to be visible cumulative effects would have to accrue. That is, if we find a "new" gene in a eukaryotic lineage, and if we assume that LGT is going on all the time during evolution, then eukaryotic genomes should become increasingly patchwork over evolutionary time, which is exactly what we see in prokaryotes: A typical bacterial or archaeal genome contains genes from all sorts of different donors (Nagies et al. 2020), and in prokaryotes, the accessory genome (that component of the genome that is constantly in flux) typically comprises about 20%–30% of an average genome, and always has, ever since the first cells roamed the ocean floor 4 billion years ago (Trost et al. 2024). In eukaryotes we see cumulative effects for differential loss, for example in the case of the microsporidian *Encephalitozoon cuniculi* (Katinka et al. 2001), reduced parasitic fungi with a 2.9 Mb genome (smaller than *Escherichia coli*) or nucleomorph genomes, eukaryotic genomes that have shrunk to <700 kb in size (Gilson 2001). But in eukaryotes, we do not see cumulative effects for LGT. How so? If eukaryotic lineages acquired just one new gene from prokaryotes per million years, on average, then after 1.5 billion years of eukaryote evolution (Mills et al. 2022) separate eukaryotic supergroups would each harbor roughly 1500 different prokaryotic genes each. If that were the case, genomes would have told us so by now. But that is not what we see. Eukaryotes have different subsets of the same ancestral collection of genes (Müller et al. 2012, Ku et al. 2015, Brueckner and Martin 2020). Reviews of eukaryote LGT (Martin 2017, Keeling 2024) tend to cover case studies of single eukaryote genes or single eukaryote genomes, that is, an odd gene here or an odd genome there. Comparative studies involving many eukaryotic lineages are still rare. Now that we have a method to estimate the frequency of last-one-out topologies, we can compare the expectation for observing such "LGT-like" topologies as a result of differential loss. The cases we tested here are fully consistent with the expectations for differential loss, alleviating the need to assume LGT involving curious mechanisms, such as gene transfer via meteorites as vectors (Bergthorsson et al. 2003) as one prominent study suggested.

A simple algorithm applied to simulated eukaryotic trees provides estimates for the frequency of last-one-out patterns resulting from a loss only model that are slightly higher than, but generally in good agreement with, observations from a recent study in which all last-one-out topologies were interpreted as evidence for LGT. Gene loss is a prevalent process in eukaryotic genome evolution. If one lineage can lose a given gene, others can as well. Gene loss can, and does, generate patterns that look just like LGT. Even for large data sets, the probability of last-one-out topologies can be surprisingly large, because, depending upon the tree, the number of losses required to account for a last-one-out topology can be small.

Author contributions

Conceptualization, W.F.M. and M.S.; methodology, N.B., W.F.M., and M.S.; investigation, N.B., W.F.M., and M.S.; writing—original draft, N.B., W.F.M., and M.S.; writing—review & editing, N.B., W.F.M., and M.S.; funding acquisition, W.F.M. and M.S.; resources, W.F.M. and M.S.; supervision, W.F.M. and M.S.

Supplementary data

Supplementary data are available at *FEMSLE Journal* online.

Conflict of interest: The authors declare that they have no competing interests.

Funding

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement number 101018894). M.S. thanks the Alexander von Humboldt Foundation for supporting his visit to Germany in 2023.

Data availability

Data files and/or online-only appendices can be found in the Dryad data repository: https://doi.org/10.5061/dryad.612jm649v. Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

References

Aguirre-Carvajal K, Cárdenas S, Munteanu CR *et al.* Rampant interkingdom horizontal gene transfer in *Pezizomycotina*? An updated inspection of anomalous phylogenies. *Int J Mol Sci* 2025;**26**:1795.

Albalat R, Cañestro C. Evolution by gene loss. *Nat Rev Genet* 2016;**17**:379–91.

Andersson SGE, Kurland CG. Reductive evolution of resident genomes. *Trends Microbiol* 1998;**6**:263–8.

Bansal MS, Alm EJ, Kellis M. Efficient algorithms for the reconciliation problem with gene duplication, horizontal transfer and loss. *Bioinformatics* 2012;**28**:i283–i291.

Bergthorsson U, Adams KL, Thomason B *et al.* Widespread horizontal transfer of mitochondrial genes in flowering plants. *Nature* 2003;**424**:197–201.

Blanc G, Wolfe KH. Functional divergence of duplicated genes formed by polyploidy during *Arabidopsis* evolution. *Plant Cell* 2004;**16**:1679–91.

Boothby TC, Tenlen JR, Smith FW *et al.* Evidence for extensive horizontal gene transfer from the draft genome of a tardigrade. *Proc Natl Acad Sci USA* 2015;**112**:15976–81.

Bremer N, Knopp M, Martin WF *et al.* Realistic gene transfer to gene duplication ratios identify different roots in the bacterial phylogeny using a tree reconciliation method. *Life* 2022;**12**:995.

Brueckner J, Martin WF. Bacterial genes outnumber archaeal genes in eukaryotic genomes. *Genome Biol Evol* 2020;**12**:282–92.

Brunet FG, Crollius HR, Paris M *et al.* Gene loss and evolutionary rates following whole-genome duplication in teleost fishes. *Mol Biol Evol* 2006;**23**:1808–16.

Burki F, Roger AJ, Brown MW *et al.* The new tree of eukaryotes. *Trends Ecol Evol* 2020;**35**:43–55.

Consortium IHGS. Initial sequencing and analysis of the human genome. *Nature* 2001;**409**:860–921.

Cote-L'Heureux A, Maurer-Alcal´a XX, Katz LA. Old genes in new places: a taxon-rich analysis of interdomain lateral gene transfer events. *PLoS Genet* 2022;**18**:e1010239.

Dagan T, Martin WF. Ancestral genome sizes specify the minimum rate of lateral gene transfer during prokaryote evolution. *Proc Natl Acad Sci USA* 2007;**104**:870–5.

Elmore MH, McGary KL, Wisecaver JH *et al.*. Clustering of two genes putatively involved in cyanate detoxification evolved recently and independently in multiple fungal lineages. *Genome Biol Evol* 2015;**7**:789–800.

Gilson PR. Nucleomorph genomes: much ado about practically nothing. *Genome Biol* 2001;**2**:reviews1022.1.

Goodman M, Czelusniak J, Moore GW *et al.* Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by cladograms constructed from globin sequences. *Syst Zool* 1979;**28**:132–63.

Gould SB, Garg SG, Handrich M *et al.* Adaptation to life on land at high O_2 via transition from ferredoxin-to NADH-dependent redox balance. *Proc Biol Sci* 2019;**286**:20191491.

Guijarro-Clarke C, Holland PWH, Paps J. Widespread patterns of gene loss in the evolution of the animal kingdom. *Nat Ecol Evol* 2020;**4**:519–23.

Happe T, Kaminski A. Differential regulation of the Fe-hydrogenase during anaerobic adaptation in the green alga *Chlamydomonas reinhardtii*. *Eur J Biochem* 2002;**269**:1022–32.

Hawkins NJ, Bass C, Dixon A *et al.* The evolutionary origins of pesticide resistance. *Biol Rev* 2019;**94**:135–55.

Hittinger C, Carrol S. Gene duplication and the adaptive evolution of a classic genetic switch. *Nature* 2007;**449**:677–81.

Hosokawa T, Kikuchi Y, Nikoh N *et al.* Strict host–symbiont cospeciation and reductive genome evolution in insect gut bacteria. *PLoS Biol* 2006;**4**:e337.

Katinka MD, Duprat S, Cornillot E *et al.* Genome sequence and gene compaction of the eukaryote parasite *Encephalitozoon cuniculi*. *Nature* 2001;**414**:450–3.

Katoh K, Misawa K, Kuma K *et al.* MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* 2002;**30**:3059–66.

Keeling PJ, Burki F. Progress towards the tree of eukaryotes. *Curr Biol* 2019;**29**:R808–17.

Keeling PJ. Horizontal gene transfer in eukaryotes: aligning theory with data. *Nat Rev Genet* 2024;**25**:416–30.

Kellis M, Birren BW, Lander ES. Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature* 2004;**428**:617–24.

Kendall DG. On the generalized 'birth-and-death' process. *Ann Math Stat* 1948;**19**:1–15.

Koutsovoulos G, Kumar S, Laetsch DR *et al.* No evidence for extensive horizontal gene transfer in the genome of the tardigrade *Hypsibius dujardini*. *Proc Natl Acad Sci USA* 2016;**113**:5053–8.

Ku C, Martin WF. A natural barrier to lateral gene transfer from prokaryotes to eukaryotes revealed from genomes: the 70% rule. *BMC Biol* 2016;**14**:89.

Ku C, Nelson-Sathi S, Roettger M *et al.* Endosymbiotic origin and differential loss of eukaryotic genes. *Nature* 2015;**524**:427–32.

Lambert A, Stadler T. Birth–death models and coalescent point processes: the shape and probability of reconstructed phylogenies. *Theor Popul Biol* 2013;**90**:113–28.

Leger MM, Gawryluk RMR. Evolution: a gene-rich mitochondrial genome sheds light on the last eukaryotic common ancestor. *Curr Biol* 2024;**34**:R776–9.

Li Y, Zhang Y, Li X *et al.* Gain-of-function mutations: an emerging advantage for cancer biology. *Trends Biochem Sci* 2019;**44**:659–74.

Martin WF, Stoebe B, Goremykin V *et al.* Gene transfer to the nucleus and the evolution of chloroplasts. *Nature* 1998;**393**:162–5.

Martin WF. Too much eukaryote LGT. *Bioessays* 2017;**39**:1700115.

Mills DB, Boyle RA, Daines SJ *et al.* Eukaryogenesis and oxygen in Earth history. *Nat Ecol Evol* 2022;**6**:520–32.

Moore CE, Archibald JM. Nucleomorph genomes. *Annu Rev Genet* 2009;**43**:251–64.

Moran NA, Bennett GM. The tiniest tiny genomes. *Annu Rev Microbiol* 2014;**68**:195–215.
Müller M, Mentel M, van Hellemond JJ *et al.* Biochemistry and evolution of anaerobic energy metabolism in eukaryotes. *Microbiol Mol Biol Rev* 2012;**76**:444–95.
Nagies FSP, Brueckner J, Tria FDK *et al.* A spectrum of verticality across genes. *PLoS Genet* 2020;**16**:1–28.
Nguyen L, Schmidt HA, von Haeseler A *et al.* IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol* 2015;**32**:268–74.
Nicholson D, Salamina M, Panek J *et al.* Adaptation to genome decay in the structure of the smallest eukaryotic ribosome. *Nat Commun* 2022;**13**:591.
Olson MV. When less is more: gene loss as an engine of evolutionary change. *Am Hum Genet* 1999;**64**:18–23.
Oshima K, Kakizawa S, Nishigawa H *et al.* Reductive evolution suggested from the complete genome sequence of a plant-pathogenic phytoplasma. *Nat Genet* 2004;**36**:27–29.
Page RDM. Maps between trees and cladistic analysis of historical associations among genes, organisms, and areas. *Syst Biol* 1994;**43**:58–77.
Richards TA, Talbot NJ. Horizontal gene transfer in osmotrophs: playing with public goods. *Nat Rev Microbiol* 2013;**11**:720–7.
Russel PE. A century of fungicide evolution. *J Agric Sci* 2005;**143**: 11–25.
Salzberg SL, White O, Peterson J *et al.* Microbial genes in the Human genome: lateral transfer or gene loss?. *Science* 2001;**292**: 1903–6.
Salzberg SL. Horizontal gene transfer is not a hallmark of the human genome. *Genome Biol* 2017;**18**:85.
Scannel DR, Byrne KP, Gordon JL *et al.* Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. *Nature* 2006;**440**:341–5.
Shen X, Opulente DA, Kominek J *et al.* Tempo and mode of genome evolution in the budding yeast subphylum. *Cell* 2018;**175**:1533–1545.e20.
Stanhope MJ, Lupas A, Italia MJ *et al.* Phylogenetic analyses do not support horizontal gene transfers from bacteria to vertebrates. *Nature* 2001;**411**:940–4.
Szöllősi GJ, Rosikiewicz W, Boussau B *et al.* Efficient exploration of the space of reconciled gene trees. *Syst Biol* 2013;**62**:901–12.
Timmis JN, Ayliffe MA, Huang CY *et al.* Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. *Nat Rev Genet* 2004;**5**:123–35.
Tovar J, Le´on-Avila G, S´anchez LB *et al.* Mitochondrial remnant organelles of Giardia function in iron-sulphur protein maturation. *Nature* 2003;**426**:172–6.
Tria F, Landan G, Dagan T. Phylogenetic rooting using minimal ancestor deviation. *Nat Ecol Evol* 2017;**1**:0193.
Trost K, Knopp MR, Wimmer JLE *et al.* A universal and constant rate of gene content change traces pangenome flux to LUCA. *FEMS Microbiol Lett* 2024;**371**:fnae068.
van de Peer Y, Maere S, Meyer A. The evolutionary significance of ancient genome duplications. *Nat Rev Genet* 2009;**10**:725–32.
van Ham RCHJ, Kamerbeek J, Palacios C *et al.* Reductive genome evolution in *Buchnera aphidicola*. *Proc Natl Acad Sci USA* 2003;**100**:581–6.
Waters E, Hohn MJ, Ahel I *et al.* The genome of *Nanoarchaeum equitans*: insights into early archaeal evolution and derived parasitism. *Proc Natl Acad Sci USA* 2003;**100**:12984–8.

Received 4 March 2025; **revised** 29 April 2025; **accepted** 20 May 2025